UNCALIBRATED STEREO VISION WITH POINTING FOR A MAN–MACHINE INTERFACE

Roberto Cipolla, Paul A. Hadfield and Nicholas J. Hollinghurst

Department of Engineering,
University of Cambridge,
Cambridge CB2 1PZ, UK.

ABSTRACT

Here we report preliminary work on a **gesture-based interface** for robot guidance. The system requires no physical contact with the operator, but uses **uncalibrated stereo vision** with **active contours** to track the position and pointing direction of a hand. With a **ground plane constraint**, it is then possible to find the indicated position in the robot's workspace, by considering only two-dimensional collineations.

The system is accurate to about 2cm in a 40cm workspace; natural operator feedback improves this to within 1cm. It is initialised by observing just 4 points on the plane.

1 INTRODUCTION

A number of systems have been proposed in the past for human–computer interaction based on hand gestures and pointing. Some systems required the user to wear a special glove or magnetic sensors [1, 2, 3]. Others using image processing have required calibration for each user's individual hand shape and posture [4].

We have developed a stereo vision pointing system as an input device for a robot manipulator, to provide a novel and convenient means for the operator to specify points for pick-and-place operations. We use *active contour* techniques [5] to track a hand in a pointing gesture, with conventional monochrome cameras and fairly modest image-processing hardware.

A single view of a pointing hand is ambiguous: its distance from the camera cannot be determined, and the slant of its orientation is hard to measure. Stereo views constrain the indicated point to a line in space, passing through the fingertip in the direction of pointing. We employ a ground-plane constraint to determine the single point on a table-top corresponding to each gesture. The use of such a constraint effectively reduces the problem to a two-dimensional one.

2 THEORY

2.1 Viewing the plane

Consider a pinhole-camera vision system viewing a plane. The viewing transformation for each camera is a plane-to-plane collineation between some world coordinate system (X, Y) and image coordinates (u, v) thus:

$$\begin{bmatrix} su \\ sv \\ s \end{bmatrix} = \mathbf{T} \begin{bmatrix} X \\ Y \\ 1 \end{bmatrix}, \qquad (1)$$

where s is a scale factor that varies for each point; and $\mathbf{T}$ is a 3×3 transformation matrix. The system is homogeneous, so we can fix $t_{33} = 1$, leaving 8 degrees of freedom. To solve for $\mathbf{T}$ we must observe at least four reference points; and, by assigning arbitrary world coordinates to these points (e.g. $(0,0)$, $(0,1)$, $(1,1)$, $(1,0)$), we define a new coordinate system on the plane, which we call *working plane coordinates*.

Now, given the image coordinates of a point anywhere in the plane, along with the image coordinates of the four reference points, it is possible to invert the relation and recover the point's working plane coordinates, which are invariant to the choice of camera location.

2.2 Pointing at the plane

With natural human pointing behaviour, there is an arbitrary distance between the hand and the indicated point: the hand is used to define a line in space, passing through the fingertip. This line will not generally be in the ground plane but intersects the plane at some point. It is that point that we aim to recover.

Let the pointing finger lie along the line l_w in space (see figure 1). Viewed by a camera, it appears to be on line l_i in the image, which is also the projection of a *plane*, $\mathcal{P}$, passing through the image line and the optical centre of the camera. This plane intersects the ground plane $\mathcal{G}$ along line l_{gp}. We know that the l_w lies in $\mathcal{P}$, and the indicated point in l_{gp}, but from one view we cannot see exactly where.

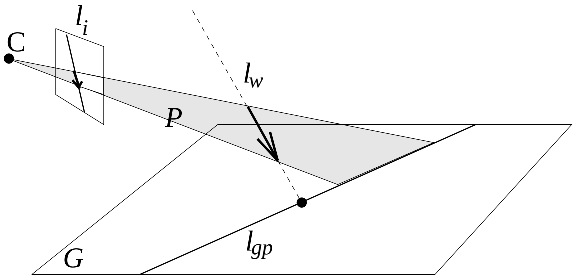

Figure 1: Projection of the finger's image line l_i onto the ground plane yields a constraint line l_{gp} on which the indicated point must lie.

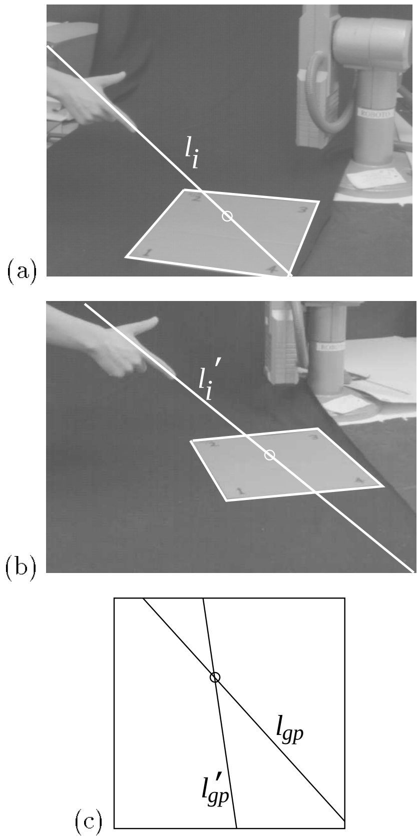

Figure 2: By taking the line of pointing in each view (a, b), transforming it into the coordinate system defined by the four reference points, and finding the intersection of the lines (c), the indicated point can be determined; this is then projected back into the images.

Note that l_i is an image of l_{gp}; that is,

$$l_i = \mathbf{T}(l_{gp}), \tag{2}$$

where $\mathbf{T}$ is the projective transformation from equation (1). If the four reference points are visible, this transformation can be inverted to find the constraint line in terms of the working plane coordinates.

Repeating this procedure with a second camera C$'$ gives us another view l_i' of the finger, and another line of constraint l_{gp}'. The two constraint lines will intersect at a point on the ground plane, which is the indicated point. Its position can now be found relative to the four reference points (figure 2).

Because all calculations are restricted to the image and ground planes, explicit 3-D reconstruction is avoided and no camera calibration is necessary. In fact, provided four points on the plane are always visible, the results will be invariant to camera motion.

3 IMPLEMENTATION

3.1 Equipment

The system is implemented on a Sun SPARCstation 10 with a Data Cell S2200 frame grabber. Images are provided by two PULNIX monochrome CCD cameras, which view the operator's hand and the working area from a distance of about 2 metres. The angle betwen the cameras is about 30°. A Scorbot ER-7 robot arm is also connected to the Sun.

3.2 Tracking the hand

We use a template-based active contour model [6] to track the extended index finger and upturned thumb of a hand in the familiar 'pointing' gesture.

Initially, the tracker deforms affinely until it has aquired the dimensions of the operator's hand; it is then restricted to rigid motions in the image plane, and a single degree of shearing (which is how the fingers appear to move unless strongly foreshortened). The pointing direction is assumed to be the orientation of the index finger; the base of the thumb is tracked merely to resolve an *aperture problem* [7] induced by the finger's long thin shape. We have deliberately avoided tracking the main part of the hand because this has a complicated shape which can vary significantly from one person to another.

The template-based tracker (figure 3) is designed to settle on a pointing hand so that one of its basis vectors is parallel to the index finger.

This makes it easy to recover the line of pointing in the image.

3.3 Pointing Experiment

Setup

For this experiment, the corners of a coloured rectangle on the tabletop are used to define the working coordinate system.

The trackers for the two cameras are initialised, one after the other, by the operator holding his hand up to a template in the image, and waiting a few seconds while it moulds itself to the contours of the finger and thumb. With both trackers running, the hand can be used as an input device by pointing to places on the tabletop.

Uncertainty

We can derive a measure of uncertainty for the trackers' position and orientation in the image by considering the "residual offsets" (offsets between actual and predicted image edges after solving for translations and deformations [6]).

From this error we calculate $\pm 2\sigma$ perturbations of the position and orientation of lines l_i and l'_i; and, by projecting these onto the ground plane, estimate the uncertainty in the position of the indicated point.

Figure 4 shows the system in operation, with an uncertainty ellipse drawn two standard deviations (95% confidence) away from the indicated point.

Performance

Humans seem to judge their own pointing direction partly by the line of sight connecting their fingertip to their *eye* rather than the finger direction itself, and this produces an offset between imagined and observed pointing directions. Subjective estimates of accuracy are in the order of 2–4cm.

Tests using an artificial pointing device (figure 5) show that our system is accurate to about 3.7% (RMS error) of the working plane coordinates, or 15mm in a 40cm workspace, which is comparable with predicted uncertainty.

3.4 Robot guidance experiment

Setup

For this experiment, the reference points are defined by observing the robot gripper as it visits 4 points in a plane (this not only defines the

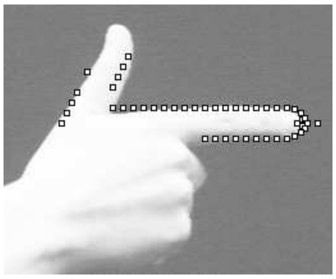

Figure 3: The template for the finger-tracking active contour.

(a) (b) (c)

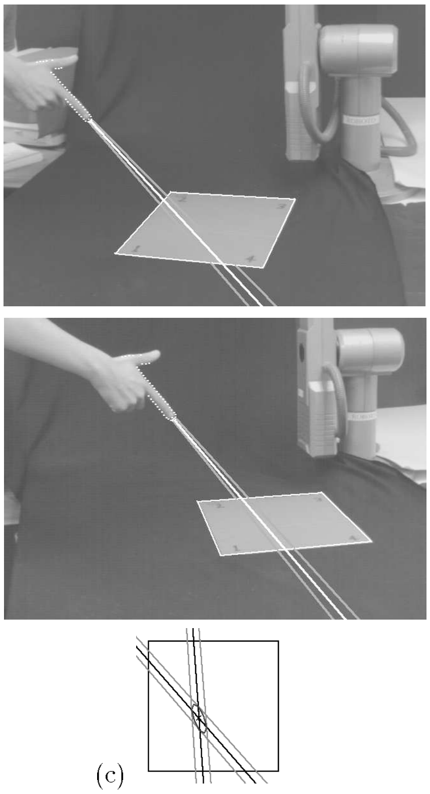

Figure 4: Uncertainty measure for the pointing lines and indicated point.

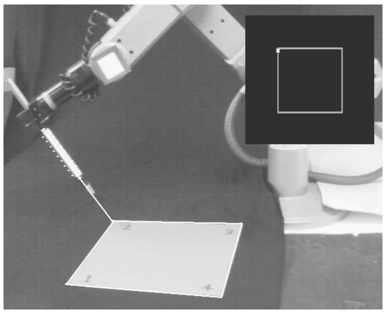

Figure 5: Mechanical pointing device used to assess the accuracy of our system

working coordinate system but relates it to the robot's own world coordinate system).

Finger-trackers operate as before, but now the robot is instructed to move repeatedly to where the hand is pointing, providing the operator with direct feedback of the system's output.

Performance

By observing feedback from the robot, the operator is able to position the gripper to within 1cm: sufficient accuracy to instruct it to pick up a small wooden block placed in its workspace (figure 6).

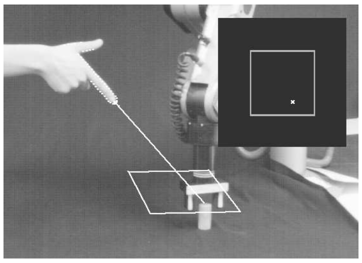

Figure 6: Gestural control of robot position. The working plane is now 50mm above the tabletop; the four reference points (white rectangle) were defined by the robot's gripper.

4 CONCLUSION

Our algorithm for resolving the pointing direction proves to be usable and stable in the presence of normal image noise. It does not require camera calibration because all calculation takes place in the image and ground planes. By tracking 4 points on the plane it could be made invariant to camera motions.

The main problem for this system is tracking a pointing hand reliably in stereo. At present, this is only possible in an environment where there is a strong constrast between the hand and the background. Tracking speed is limited by our hardware (a single Sun SPARCstation) and could be improved by adding purpose-built image processing equipment.

When a human operator is included in a system, they can, almost subconsciously, feed back on the output and compensate for small systematic errors. In this manner, a human can indicate positions with sufficient accuracy to guide pick-and-place operations.

ACKNOWLEDGEMENTS

We are indebted to Mr. Masaaki Fukumoto and Dr. Yasuhito Suenaga of the NTT Human Interface Laboratories, Yokosuka, Japan, who collaborated with Roberto Cipolla on an earlier implementation of this algorithm.

We gratefully acknowledge the financial support of EPSRC, and the donation of a robot manipulator by the Olivetti Research Laboratory, Cambridge.

References

[1] R. A. Bolt. 'Put-that-there': Voice and gesture at the graphics interface. *ACM-SIGGRAPH*, vol. 14 no. 3, pp 262–270, 1980.

[2] D. Wiemer and S. G. Ganapathy. A synthetic visual environment with hand gesturing and voice input. *Proc. CHI'89*, pp 235-240, 1989.

[3] R. Cipolla, Y. Okamoto and Y. Kuno. Qualitative visual interpretation of 3D hand gestures using motion parallax. *Proc. IAPR Workshop on Machine Vision Applications*, pp 477–482, Tokyo, 1992.

[4] M. Fukumoto, K. Mase and Y. Suenaga. Realtime detection of pointing actions for a glove-free interface. *Proc. IAPR Workshop on Machine Vision Applications*, pp 473–476, Tokyo, 1992.

[5] R. Cipolla and A. Blake. Surface shape from the deformation of apparrent contours. *Int. J. Computer Vision*, vol. 9 no. 2 pp 83–112, 1992.

[6] N. J. Hollinghurst and R. Cipolla. Uncalibrated stereo hand–eye coordination. *Image and Vision Computing*, vol. 12 no. 3 pp 187–192, 1994.

[7] S. Ullman. *The interpretation of visual motion.* MIT Press, Cambridge, Ma., USA. 1979.

NOTE: This paper will appear in the proceedings of the IAPR workshop on machine vision applications MVA'94. December 1994, Tokyo.